The Battle of the Neighborhoods – Report

Introduction and Business Problem

The City of New York, usually called either New York City (NYC) or simply New York (NY), is the most populous city in the United States. With an estimated 2018 population of 8,398,748 distributed over a land area of about 302.6 square miles (784 km2), New York is also the most densely populated major city in the United States.[Ref : https://en.wikipedia.org/wiki/New_York_City]

Being a global hub of business and commerce, city has a major centers for banking and finance, retailing, world trade, transportation, tourism, etc.

This also means that the market is highly competitive. As it is highly developed city so cost of doing business is also one of the highest. Thus, any new business venture or expansion needs to be analyzed carefully. Proper analysis of data would not only reduce risks of entering new market, it will also help arrive at the Return on Investment.

Problem Description

With more than 8 million people in New York City, you'd expect some great chow and some culinary diversity – and that's exactly what you get. The city manages to bring dishes from all over the world and make them its own.

Much of the cuisine usually associated with New York City stems in part from its large community of Ashkenazi Jews and their descendants.

Like the Askenazi-Jewish community, much of the cuisine usually associated with New York City stems in part from its large community of Italian-Americans and their descendants. Much of New York City's Italian fare has become popular around the globe, especially New York-style pizza.

Chino-Latino cuisine associated with New York City stems, by and large, to the earliest migration of Chinese migrants to Cuba in the mid-1800s.

Mobile food vendors - Some 4,000 licensed by the city

The occurrence of the Cuban and Chino fuse, had been established when the Chinese began to migrate to Cuba had been viewed as a unique immersion of a new and diverse part of culture.

So it is evident that to survive in such competitive market, metaculous planning is a must to penetrate this market. Various factors need to be studied in order to decide on the Location such as:

1. New York Population
2. New York City Demographics
3. Are there any Vegetable/fruits market, Farmers Markets, Wholesale markets etc. nearby so that the ingredients can be purchased fresh to maintain quality and cost?
4. Are there any venues like Gyms, Entertainment zones, Parks etc. nearby where floating population is high etc.
5. Who are the competitors in that location?
6. Cuisine served / Menu of the competitors
7. Segmentation of the Borough
8. Untapped markets
9. Saturated markets etc.

The list can go on...

A well-funded XYZ Company Ltd. needs to choose precise location to start its first venture. Deciding the first location is crucial since, if this is successful they can replicating the same in other locations would make sense.

Target Audience:

To recommend the correct location, XYZ Company Ltd has appointed me to lead of the Data Science team. The objective is to locate and recommend to the management which neighborhood of New York city will be best choice to start a restaurant. The Management also expects to understand the rationale of the recommendations made.

This would interest anyone who wants to start a new restaurant in New York city.

Success Criteria:

The success criteria of the project will be a good recommendation of borough/Neighborhood choice to XYZ Company Ltd based on Lack of such restaurants in that location and nearest suppliers of ingredients.

Data:

To solve the problem, we will need the following data:

- List of neighbourhood of New York City.
- Latitude and longitude coordinates of those neighbourhoods.
- Data related to farmers market, city population, demographics and cuisine. We will use this data to perform clustering on the neighbourhoods.

The city analyzed in this project is New York City. Following datasets will be used for analyzing New York City.

Data 1 Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segment these neighborhoods and explore them, we need dataset which contains 5 boroughs and neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood.

This dataset exists for free on the web. Link to the dataset is :
https://geo.nyu.edu/catalog/nyu_2451_34572

Data 2 Second dataset used is the DOHMH Farmers Markets and Food Boxes dataset. In this we will use the data of Farmers Markets.

https://data.world/city-of-ny/8vwk-6iz2/workspace/file?filename=dohmh-farmers-markets-1.csv

Website-https://data.world/city-of-ny/8vwk-6iz2

GrowNYC's Fresh *Food Box* Program is a food access initiative that enables under-served communities to purchase fresh, healthy, and primarily regionally grown produce well below traditional retail prices.

A *farmers' market* is often defined as a public site used by two or more local or regional producers for the direct sale of farm products to consumers. In addition to fresh fruits and vegetables, markets may sell dairy products, fish, meat, baked goods, and other minimally processed foods.

Data 3 For the below analysis we will get data from Wikipedia as given below:

- New York Population
- New York City Demographics
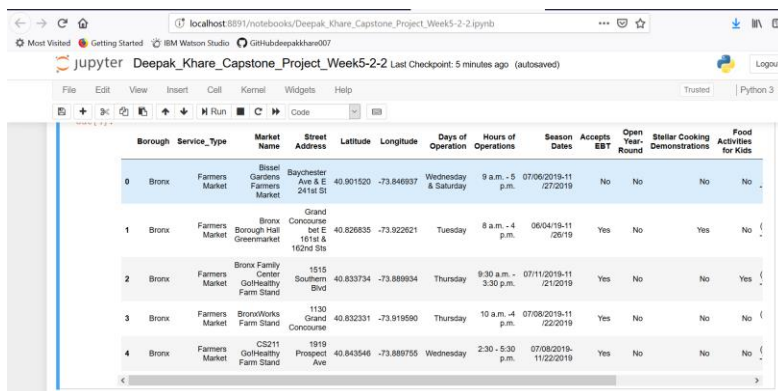- Cuisine of New York city

https://en.wikipedia.org/wiki/New_York_City
https://en.wikipedia.org/wiki/Economy_of_New_York_City
https://en.wikipedia.org/wiki/Portal:New_York_City
https://en.wikipedia.org/wiki/Cuisine_of_New_York_City
https://en.wikipedia.org/wiki/List_of_Michelin_starred_restaurants_in_New_York_City

Data 4 New York City geographical coordinates data will be utilized as input for the Foursquare API, that will be leveraged to provision venues information for each neighborhood. We will use the Foursquare API to explore neighborhoods in New York City. The below is image of the Foursquare API data.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896521	-73.844680	Pharmacy
2	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898283	-73.850478	Caribbean Restaurant
3	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
4	Wakefield	40.894705	-73.847201	Dunkin Donuts	40.890631	-73.849027	Donut Shop

Business Understanding:

Our main goal is to get optimum location for new restaurant business in New York City for XYZ Company.

Analytic Approach:

New York City neighbourhood has a total of 5 boroughs and 306 neighborhoods.

The project is segmented in two parts –

1. Clustering of Manhattan and Brooklyn
2. Clustering of Bronx, Queens and Staten Island.

This is done with the help of following Exploratory data analysis.

Exploratory Data Analysis:

Data 1- New York City Geographical Coordinates Data.

1. In this we load the data and explore data from newyork_data.json file.
2. Transform the data of nested python dictionaries into a pandas dataframe.
3. This dataframe contains the geographical coordinates of New York City neighborhoods.
4. This data will used to get Venues data from Foursquare.
5. We used geopy and folium libraries to create a map of New York city with neighborhoods superimposed on top.
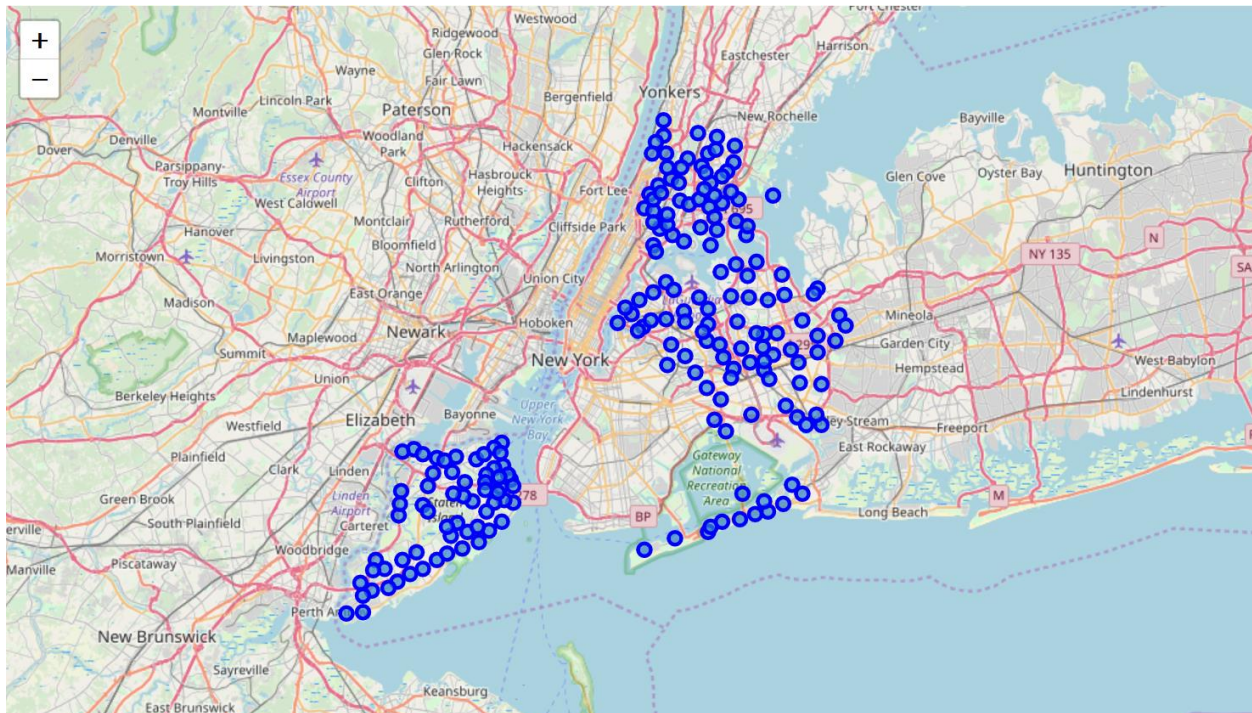
Data 2- DOHMH Farmers Markets and Food Boxes dataset.

In this we will be using the data of Farmers Markets data. There are a total of 122 Farmers. Markets in New York City. Highest number are in Manhattan and Brooklyn. And lowest in Queens, Bronx and Staten Island. The proof of this is given below.
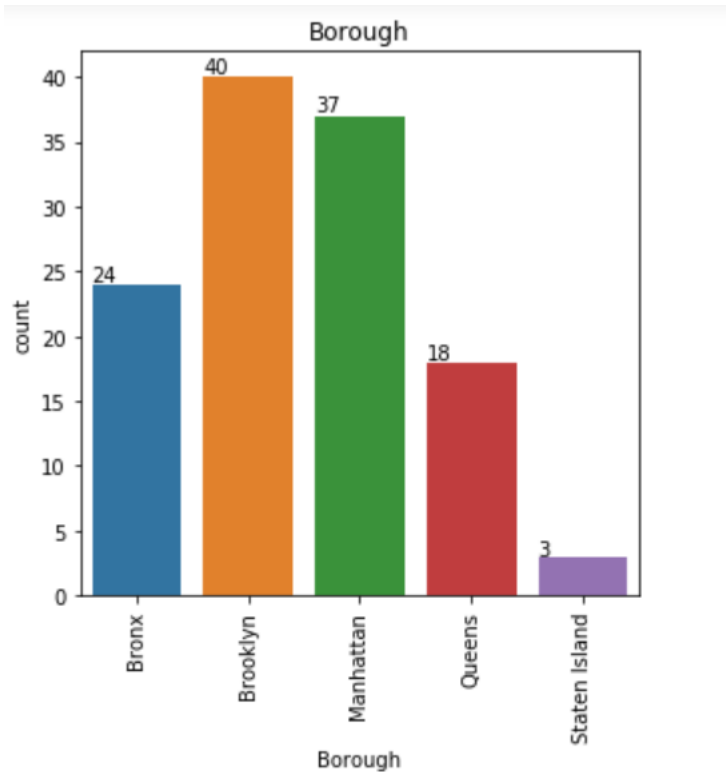
Manhattan and Brooklyn has highest numbers of Farmers Markets. We used geopy and folium libraries to create a map to visualize farmers markets of New York City.

Farmers Market visualization-New York City
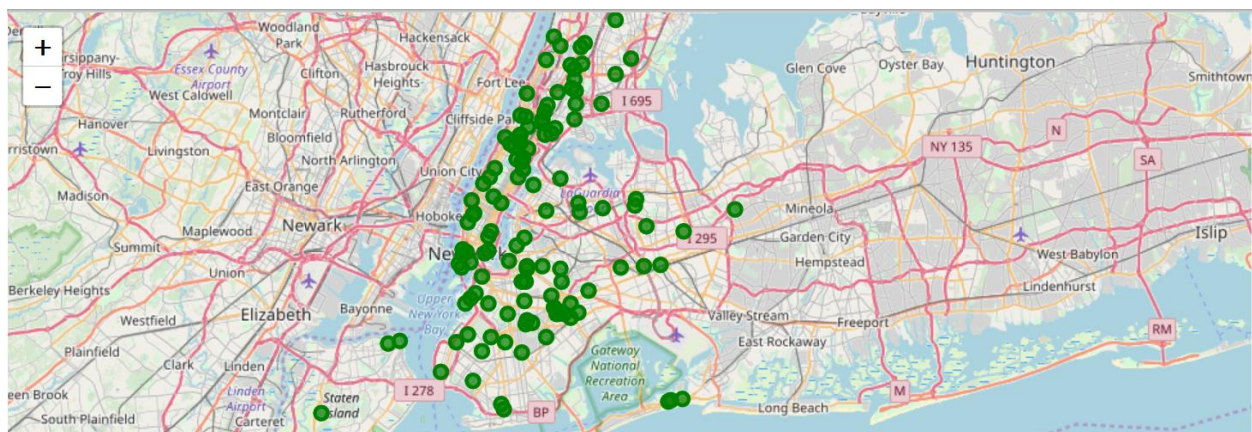
Data 3: To analyze New York City Population, Demographics and Cuisine, scrapped the data from Wikipedia pages as given above in the data section.

1. New York Population: Insights from the Data: Manhattan (New York County) is the geographically smallest and most densely populated borough.
2. Manhattan's (New York County's) population density of 72,033 people per square mile (27,812/km²) in 2015 makes it the highest of any county in the United States and higher than the density of any individual American city.
3. Brooklyn (Kings County), on the western tip of Long Island, is the city's most populous borough. Queens (Queens County), on Long Island north and east of Brooklyn, is geographically the largest borough.
4. Queens (Queens County), on Long Island north and east of Brooklyn, is geographically the largest borough.

Cuisine of New York City:

This data has been manually prepared. Data is taken from

Wikipedia page - https://en.wikipedia.org/wiki/Cuisine_of_New_York_City. Using this data we did word cloud.

NEW YORK CITY CUISINE: Most Preferred Food cuisines in New York City are Italian, Purto Rican, Mexican, Jewish, Indian, Pakistani & Dominican.

Data 4: New York City geographical coordinates data has been utilized as input from the Foursquare API, that has been leveraged to provision venues information for each neighborhood.

We used the Foursquare API data to explore neighborhoods in New York City.

Brooklyn and Manhattan Visualization:

```
In [17]: colnames = ['Neighborhood', 'Neighborhood Latitude', 'Neighborhood Longitude', 'Venue', 'Venue Latitude', 'Venue Longitu
         BM_venues = pd.read_csv('BM_venues.csv', skiprows=1, names=colnames)
         BM_venues.columns = BM_venues.columns.str.replace(' ', '')
         BM_venues.head()
```

Out[17]:

	Neighborhood	NeighborhoodLatitude	NeighborhoodLongitude	Venue	VenueLatitude	VenueLongitude	VenueCategory
0	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
1	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Sam's Pizza	40.879435	-73.905859	Pizza Place
4	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop

```
In [18]: BM_venues.shape
```

Out[18]: (9748, 7)

Visualize the BM_Venues data

```
In [19]: def Venues_Map(Borough_name, Borough_neighborhoods):

             # Use geopy library to get the latitude and longitude values
             geolocator = Nominatim(user_agent="Jupyter")
             Borough_location = geolocator.geocode(Borough_name) #'Brooklyn, NY'
             Borough_latitude = Borough_location.latitude
             Borough_longitude = Borough_location.longitude
             print('The geographical coordinates of "{}" are {}, {}.'.format(Borough_name, Borough_latitude, Borough_longitude))
```
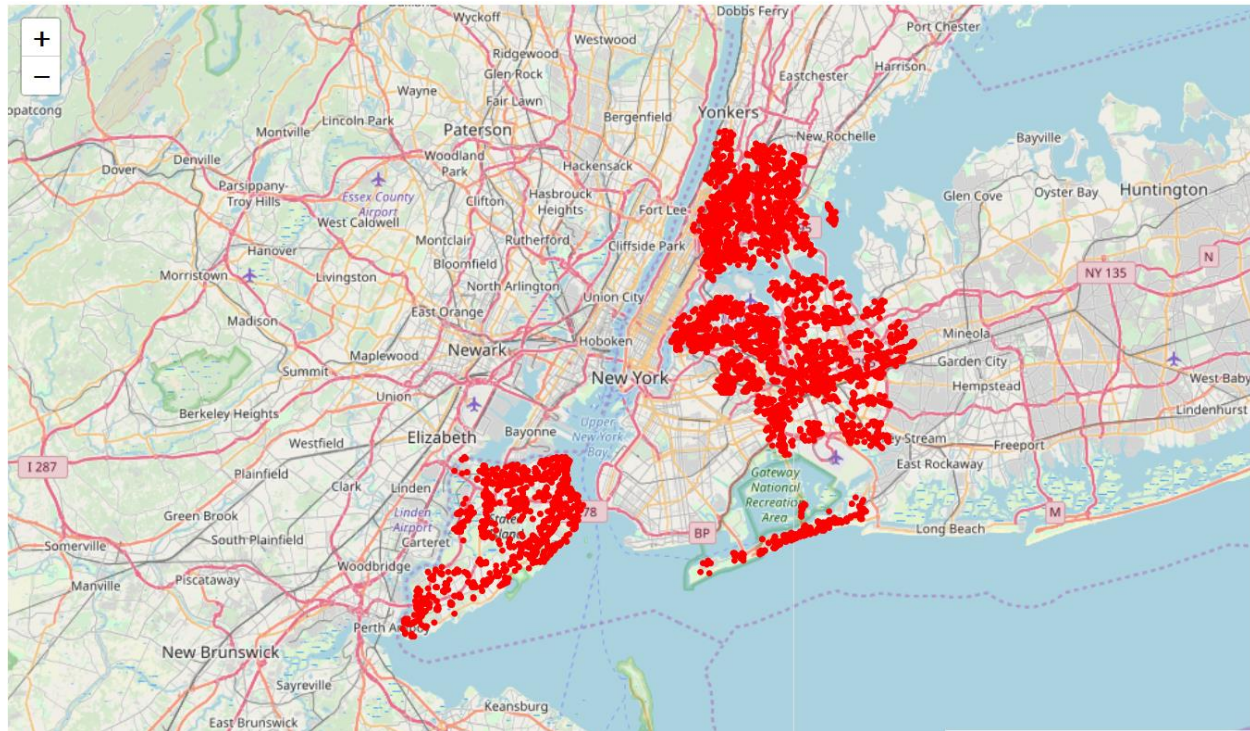
Bronx, Queens and Staten Island Venues Visualization:

Out[47]:

	Neighborhood	NeighborhoodLatitude	NeighborhoodLongitude	Venue	VenueLatitude	VenueLongitude	VenueCategory
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Ripe Kitchen & Bar	40.898152	-73.838875	Caribbean Restaurant
2	Wakefield	40.894705	-73.847201	Ali's Roti Shop	40.894036	-73.856935	Caribbean Restaurant
3	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
4	Wakefield	40.894705	-73.847201	Jackie's West Indian Bakery	40.889283	-73.843310	Caribbean Restaurant

Visualize the BQS_Venues data

```
In [48]: Venues_Map('New York City, NY', BQS_venues)
```

```
The geographical coordinates of "New York City, NY" are 40.7127281, -74.0060152.
The "New York City, NY" dataframe has 387 different venue types and 194 neighborhoods.
```

RESULTS:

From this venues data we used only the restaurant data for Brooklyn & Manhattan clustering as well as Bronx, Queens and Staten Island clustering.

For this project, we have focused only on the restaurant business.

K-Means clustering algorithm based on mean occurrence of venue category:

K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. It uses iterative refinement approach.

Brooklyn and Manhattan:

In the below Map Visualization, we can see the different types of clusters created by using K-Means for Brooklyn and Manhattan.
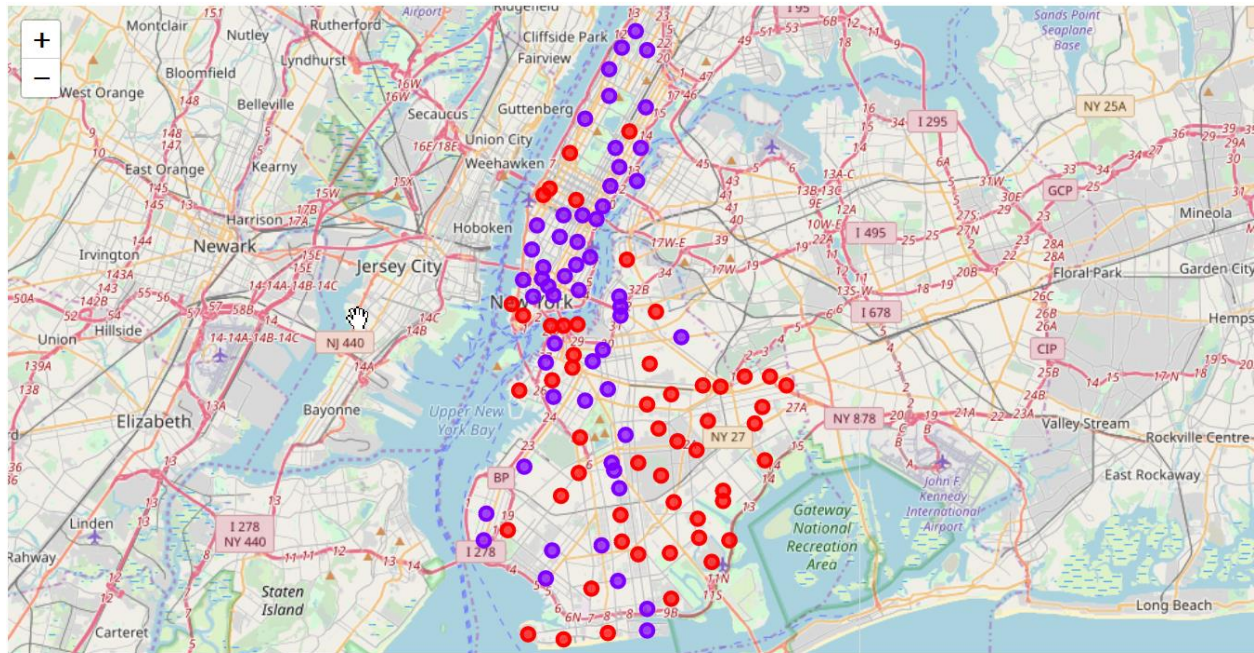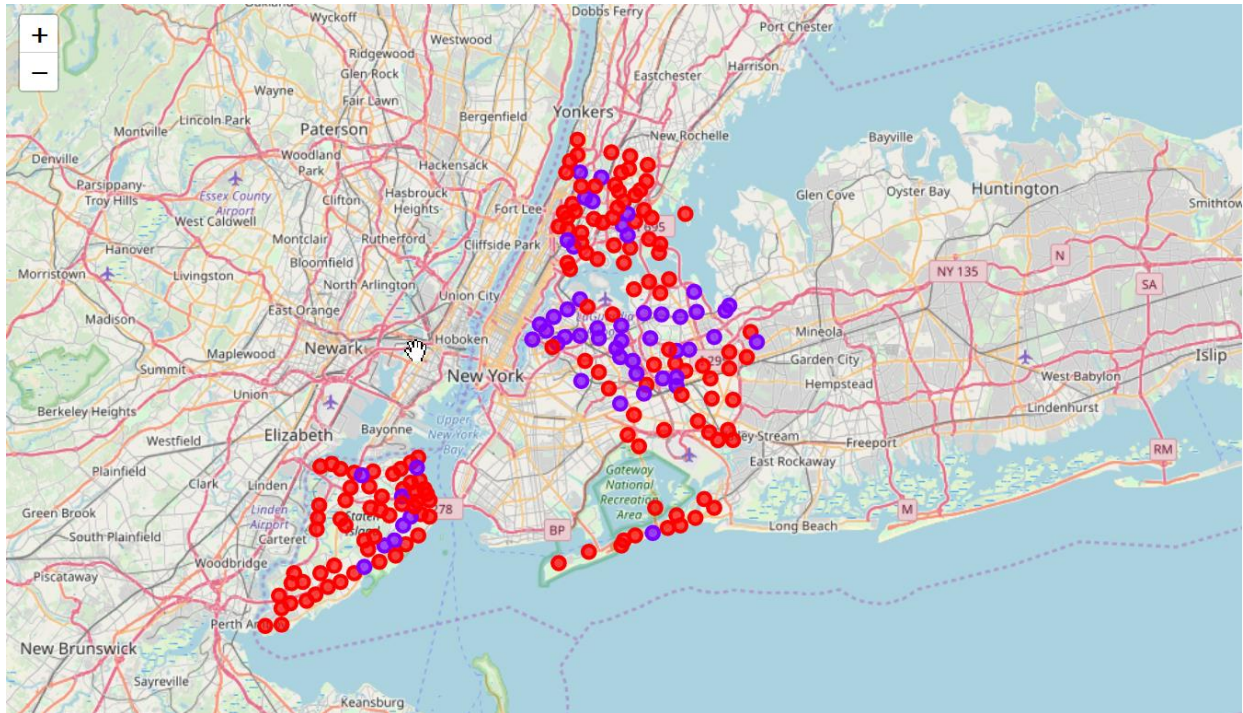

Cluster 0: The Total and Total Sum of cluster0 has smallest value. It shows that the market is not saturated.

Cluster1: The Total and Total Sum of cluster 1 has the highest value. It shows that the markets are saturated. Number of restaurants are very high. There are no untapped neighborhoods in Brooklyn and Manhattan.

Bronx, Queens and Staten Island:

In the below Map Visualization, we can see the different types of clusters created by using K-Means for Bronx, Queens and Staten Island.

Cluster 0:

The Total and the Total Sum of cluster 0 has smallest value. It shows that the market is not saturated. There are untapped neighborhoods.

Cluster1:

The Total and the Total Sum of cluster 1 has the highest value. It shows that the markets are saturated. Number of restaurants are very high.

DISCUSSION:

1. There is scope to increase Farmers markets in Bronx, Queens and Staten Island.
2. There is scope to explore cuisines of various countries in Bronx, Queens and Staten Island.
3. In Manhattan and Brooklyn restaurants of cuisines of many countries are available. So if risk can be taken with great menu on board. It also shows people love eating cuisines of various countries.

CONCLUSION:

Analysis is performed on limited data. So the end result may be right or can be away from the best result. If there are a lot of restaurants probably there is a lot of demand.

Brooklyn and Manhattan has high concentration of restaurant business. Very competitive market.

Bronx, Queens and Staten Island also has a good number of restaurants but not as many as required. So this can be explored. As per the neighbourhood or restaurant type mentioned like Italian or Indian Restaurant analysis can be done. A venue with lowest risk and competition can be identified.